Thinking Statistically about Analyzing Global Environmental Datasets

By Noel Cressie

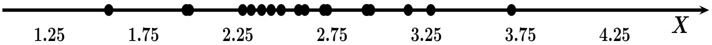
Director, Program in Spatial Statistics and Environmental Sciences
The Ohio State University
(http://www.stat.osu.edu/∼sses)

Outline

Thinking Statistically

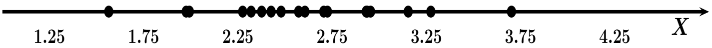
- Data are $\{X_1, \ldots, X_n\}$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ on a given day: Summarize with a scatter diagram and $(\overline{X}, S_X^2)$
  Example: X = aerosol
- Data are $\{Y_1, \ldots, Y_n\}$ at *same* locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ on the *same* day: Summarize with a scatter diagram and $(\overline{Y}, S_Y^2)$
  Example: Y = precipitation

Summaries

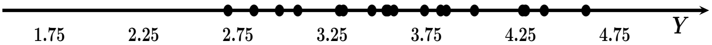


$$\overline{X} = 2.6337 \quad S_X^2 = 0.2623$$

Summaries



$$\overline{X} = 2.6337 \quad S_X^2 = 0.2623$$



$$\overline{Y} = 3.6483 \quad S_Y^2 = 0.2976$$

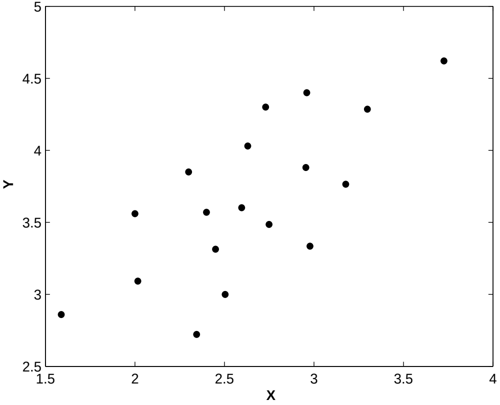
Regression

Is Y related to X?

Data are $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ on a given day: Summarize with a scatter plot and a statistical regression relationship,

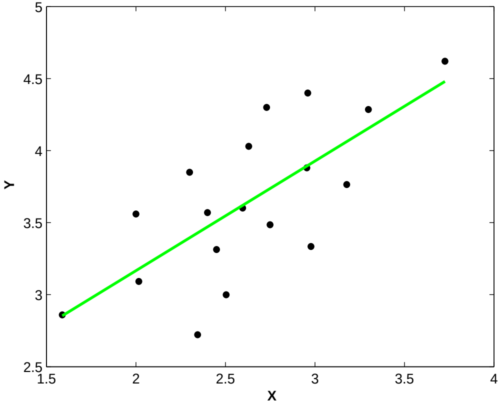
$$Y = \beta_0 + \beta_1 X + error$$

Scatterplot of Y versus X



$$Y = \beta_0 + \beta_1 X + error$$

Regression of Y on X



$$Y = \beta_0 + \beta_1 X + error$$

Regression and Statistical Inference

- The fitted line $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X$ "fills in" missing values of Y for *any* given value of X.
- Try to imagine science without regression! We would have to let X vary on a very fine scale and measure each Y associated with these many X's. Given $X = X_0$, we *infer* $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$.
- Are we certain that $\widehat{Y}_0 = Y_0$, the true value? No, but we are 95% certain that

$$Y_0 \in (\widehat{Y}_0 - 1.96\sigma_0\,,\ \ \widehat{Y}_0 + 1.96\sigma_0)\,,$$

where $\sigma_0^2 = E(\widehat{Y}_0 - Y_0)^2 = (\text{well known formula})$

Regression and Statistical Inference

- The fitted line $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X$ "fills in" missing values of Y for *any* given value of X.
- Try to imagine science without regression! We would have to let X vary on a very fine scale and measure each Y associated with these many X's. Given $X = X_0$, we *infer* $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$.
- Are we certain that $\widehat{Y}_0 = Y_0$, the true value? No, but we are 95% certain that

$$Y_0 \in (\widehat{Y}_0 - 1.96\sigma_0\,,\ \ \widehat{Y}_0 + 1.96\sigma_0)\,,$$

where $\sigma_0^2 = E(\widehat{Y}_0 - Y_0)^2 = (\text{well known formula})$

Regression and Statistical Inference

- The fitted line $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X$ "fills in" missing values of Y for *any* given value of X.
- Try to imagine science without regression! We would have to let X vary on a very fine scale and measure each Y associated with these many X's. Given $X = X_0$, we *infer* $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$.
- Are we certain that $\widehat{Y}_0 = Y_0$, the true value? No, but we are 95% certain that

$$Y_0 \in (\widehat{Y}_0 - 1.96\sigma_0\,,\;\; \widehat{Y}_0 + 1.96\sigma_0)\,,$$

where $\sigma_0^2 = E(\widehat{Y}_0 - Y_0)^2 = (\text{well known formula})$

Spatial Regression

- Y_0 = aerosol at location $\mathbf{s}_0$
- $\{X_j\}$ = aerosol values at *nearby* locations $\{\mathbf{s}_j\}$
- Multiple regression:

$$Y_0 = \lambda_0 + \lambda_1 X_1 + \cdots + \lambda_m X_m + \textit{ error}$$

- Fill in missing aerosol values Y_0 from observations $\{X_j\}$:

$$\widehat{Y}_0 = \widehat{\lambda}_0 + \widehat{\lambda}_1 X_1 + \cdots + \widehat{\lambda}_m X_m$$

and

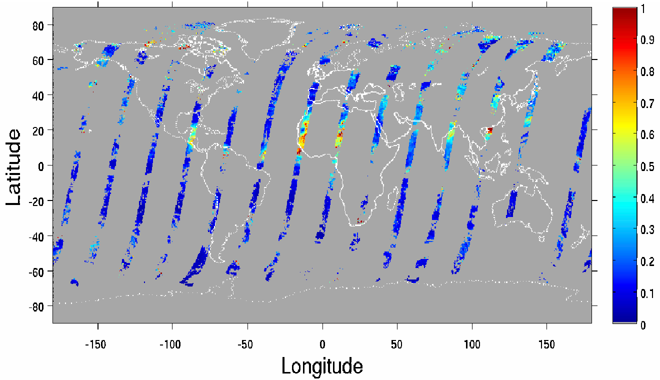
$$\sigma_0^2 = E(\widehat{Y}_0 - Y_0)^2$$

Recall

$$Y_0 \in (\widehat{Y}_0 - 1.96\sigma_0, \widehat{Y}_0 + 1.96\sigma_0)$$

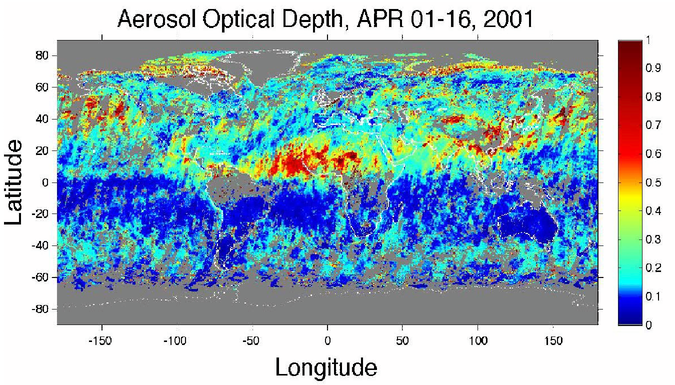
with 95% probability

MISR Level 2 Aerosol Optical Depth (AOD) Coverage on April 1, 2002



"Non-retrieval" occurs when radiance data are missing, clouds are present, or the algorithm fails

MISR Level AOD (band 3, 558 nm) 0.5×0.5 degree on the globe



720×360 pixels in a half-degree by half-degree global map; Level 2 data are averaged within each pixel

Spatial Statistical Model

- Observed variable

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s})\,; \quad \mathbf{s} \in D\,,$$

  where $\varepsilon(\cdot)$ is uncorrelated measurement error

- Hidden process with p regressors $t_1(\mathbf{s}), \ldots, t_p(\mathbf{s})$

$$Y(\mathbf{s}) = \mathbf{T}(\mathbf{s})'\boldsymbol{\alpha} + \nu(\mathbf{s})\,; \quad \mathbf{s} \in D\,,$$

  where $\mathbf{T}(\mathbf{s}) \equiv (t_1(\mathbf{s}), \ldots, t_p(\mathbf{s}))'$, regression parameters $\boldsymbol{\alpha}$ are fixed but unknown, and $\nu(\cdot)$ has mean zero

- Variance-covariance structure

$$\mathrm{var}(\varepsilon(\mathbf{s})) = \sigma^2 v(\mathbf{s})\,, \quad \mathrm{cov}(\nu(\mathbf{s}), \nu(\mathbf{t})) = C(\mathbf{s}, \mathbf{t})$$

Covariance Matrix of Data

Denote

$$\begin{aligned} \mathbf{V} &\equiv \mathrm{diag}(v(\mathbf{s}_1), \ldots, v(\mathbf{s}_n)) \\ \mathbf{C} &\equiv \{C(\mathbf{s}_i, \mathbf{s}_j)\} \end{aligned}$$

Define

$$\boldsymbol{\Sigma} \equiv \mathrm{var}((Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))')$$

Then

$$\boldsymbol{\Sigma} = \mathbf{C} + \sigma^2 \mathbf{V}$$

Kriging Equations

- The kriging predictor at location $\mathbf{s}_0$ is

$$\begin{aligned}\widehat{Y}(\mathbf{s}_0) &= \mathbf{T}(\mathbf{s}_0)'\widehat{\alpha} + \mathbf{k}(\mathbf{s}_0)'(\mathbf{Z} - \mathbf{T}\widehat{\alpha}) \\ &= \widehat{\lambda}_1 Z(\mathbf{s}_i) + \cdots + \widehat{\lambda}_n Z(\mathbf{s}_n)\end{aligned}$$

with

$$\mathbf{T} \equiv (\mathbf{T}(\mathbf{s}_1), \ldots, \mathbf{T}(\mathbf{s}_n))', \quad \widehat{\alpha} = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{Z},$$

$$\text{and} \quad \mathbf{k}(\mathbf{s}_0)' = \mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}$$

- Now let $\mathbf{s}_0$ vary to create a map of $\widehat{Y}(\cdot)$

Kriging Equations, ctd.

- The root-mean-squared prediction error (kriging standard error) of $\widehat{Y}(\mathbf{s}_0)$ is:

$$\begin{aligned}\sigma_k(\mathbf{s}_0) \quad = \quad & \{\mathbf{C}(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{k}(\mathbf{s}_0)'\boldsymbol{\Sigma}\mathbf{k}(\mathbf{s}_0) \\ & +(\mathbf{T}(\mathbf{s}_0) - \mathbf{T}'\mathbf{k}(\mathbf{s}_0))'(\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1} \\ & (\mathbf{T}(\mathbf{s}_0) - \mathbf{T}'\mathbf{k}(\mathbf{s}_0))\}^{1/2}\end{aligned}$$

- $Y(\mathbf{s}_0) \in (\widehat{Y}(\mathbf{s}_0) - 1.96\sigma_k(\mathbf{s}_0), \widehat{Y}(\mathbf{s}_0) + 1.96\sigma_k(\mathbf{s}_0))$ with 95% probability.

Kriging of Very Large Data Sets

- The bottleneck for kriging computations is inverting the $n \times n$ covariance matrix $\mathbf{\Sigma}$
- Directly inverting $\mathbf{\Sigma}$ becomes very slow, even infeasible, when sample size n is large (e.g., when $n > 3,000$)
- For the MISR Level 3 global AOD data on half-degree resolution, n could be as much as $720 \times 360 = 295,200$

Fixed Rank Kriging (FRK)

- $\mathbf{\Sigma} = \mathbf{C} + \sigma^2\mathbf{V}$
- Assume $\mathbf{C}_{n\times n}$ is of the form

$$\mathbf{C} = \mathbf{S}_{n\times r}\mathbf{K}_{r\times r}\mathbf{S}'_{r\times n}\,,$$

  with r fixed basis functions,
  $\{S_j(\cdot) : j = 1, \ldots, r\} \equiv \mathbf{S}(\cdot)$, and $r << n$
- We have derived explicit forms of $\mathbf{\Sigma}^{-1}$, and hence of the kriging solutions, when n is very large. Inversion of $r \times r$ matrices and $n \times n$ diagonal matrices are involved
- References: Cressie and Johannesson (2006, 2008) and Shi and Cressie (2007)

Fixed Rank Kriging Equations

$$\boldsymbol{\Sigma}^{-1} = (\sigma^2\mathbf{V})^{-1} - (\sigma^2\mathbf{V})^{-1}\mathbf{S}(\mathbf{K}^{-1} + \mathbf{S}'(\sigma^2\mathbf{V})^{-1}\mathbf{S})^{-1}\mathbf{S}'(\sigma^2\mathbf{V})^{-1}$$

$$\begin{aligned}
\widehat{Y}(\mathbf{s}_0) &= \mathbf{T}(\mathbf{s}_0)'\widehat{\boldsymbol{\alpha}} + \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{T}\widehat{\boldsymbol{\alpha}}) \\
\widehat{\boldsymbol{\alpha}} &= (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}
\end{aligned}$$

$$\begin{aligned}
\sigma_k(\mathbf{s}_0) = \; & \{\mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}(\mathbf{s}_0) - \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}'\mathbf{K}\mathbf{S}(\mathbf{s}_0) \\
& + (\mathbf{T}(\mathbf{s}_0) - \mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s}_0))'(\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1} \\
& \quad (\mathbf{T}(\mathbf{s}_0) - \mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s}_0))\}^{1/2}
\end{aligned}$$

Estimating $\mathbf{K}$

- Given the empirical, binned covariance matrix $\widehat{\boldsymbol{\Sigma}}_M$, we find $\widehat{\mathbf{K}}$ and $\widehat{\sigma}^2$ that minimize the Frobenius norm between $(\mathbf{S}_M \mathbf{K} \mathbf{S}_M' + \sigma^2 \mathbf{V}_M)$ and $\widehat{\boldsymbol{\Sigma}}_M$
- Frobenius norm between two matrices:

$$\|\mathbf{A} - \mathbf{B}\|^2 = tr((\mathbf{A} - \mathbf{B})'(\mathbf{A} - \mathbf{B}))$$

(We use a weighted version of the Frobenius norm to estimate $\mathbf{K}$ and σ^2)

Estimating **K**, ctd.

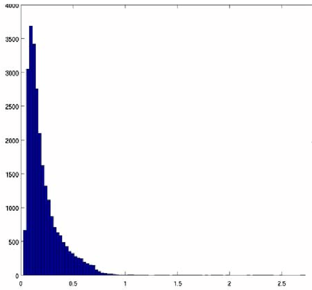
- By minimizing the Frobenius norm, we obtain

$$\widehat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'(\widehat{\boldsymbol{\Sigma}}_M - \widehat{\sigma}^2\mathbf{V}_M)\mathbf{Q}(\mathbf{R}^{-1})' \,,$$

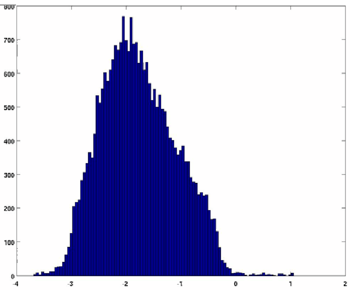
  where **Q** and **R** are obtained from the *Q*-*R* decomposition of $\mathbf{S}_M$
- Now we have all we need for FRK!

MISR Level 3 AOD Data, APR 01-16, 2001

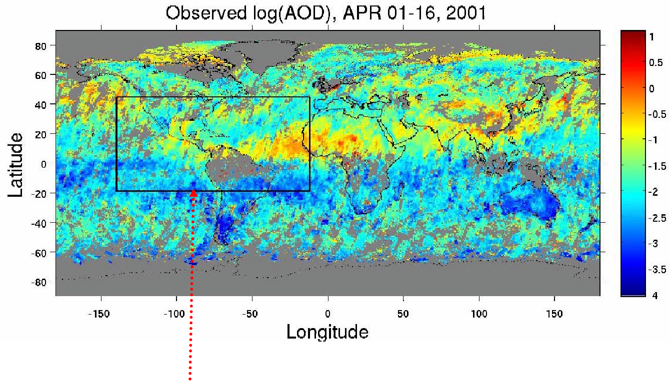
Histogram of AOD



Histogram of log(AOD)



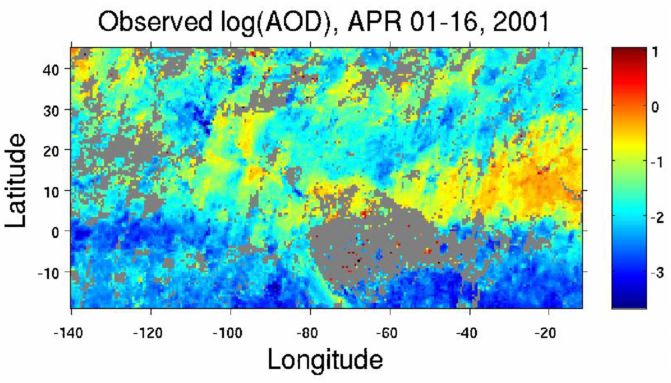
$Z(\mathbf{s}) = \log(AOD(\mathbf{s}))$



Region of Interest (D): $\mathrm{lat}(-20, 45)$, $\mathrm{lon}(-140, 12)$

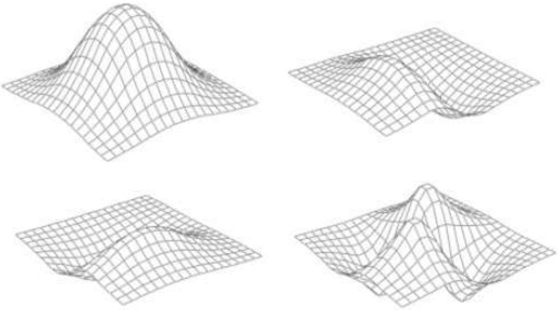
Region of Interest

Total number of pixels: $32,768$ (128×256)
Pixels observed: $n = 25,897$ (79% of total no. pixels)



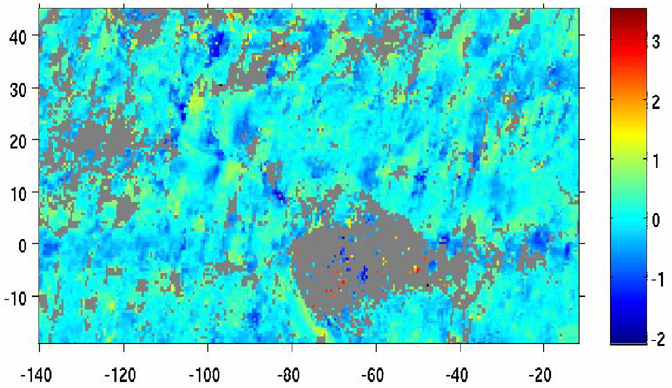
Basis Functions

- $\mathbf{T}(\cdot)$ and $\mathbf{S}(\cdot)$ are chosen as two-dimensional W-wavelet basis functions (Kwong and Tang, 1995; Nychka et al., 2002). They are tensor products of one-dimensional W-wavelets



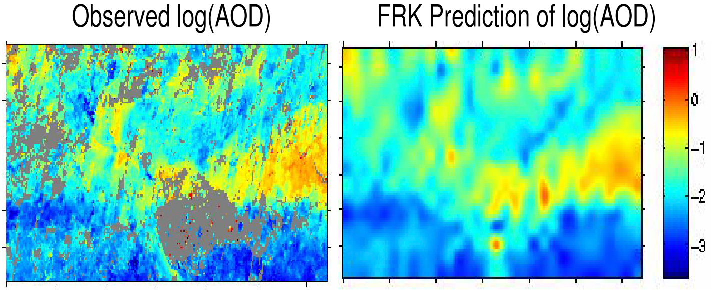
Residuals from OLS

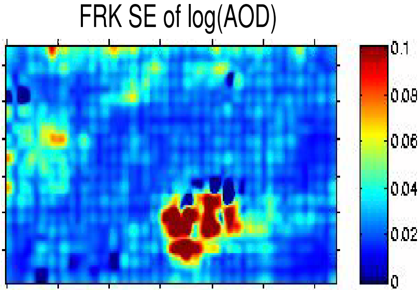
$$Z_R(\mathbf{s}_0) \equiv Z(\mathbf{s}_0) - \mathbf{t}(\mathbf{s}_0)'\widehat{\alpha}_{OLS}$$



These residuals are used to fit **K** in FRK

FRK Results





Components of $\mathbf{S}(\cdot)$ are W-wavelets at finer resolutions

Fixed rank of $\mathbf{K} = r =$ no. of basis functions $= 300$

Computation Time

Computation is done in Matlab on a Pentium 4 dual core 3.0Ghz, 2GB memory Linux Machine. Time unit: seconds

p	r	Fit $\mathbf{K}$, σ^2	Prediction
32	300	1895.3	156.0
32	400	2536.8	278.7
64	300	1894.4	170.2
64	400	2536.6	287.6

FRK Today

- FRK: Computations are $O(n)$ as compared to the usual kriging algorithms, where they are $O(n^3)$. *All n* data are used to produce the optimal linear predictor
- Multivariate: The same ideas could be used to combine MISR and MODIS data to obtain an optimal predictor of AOD (data fusion)
- Space-time: Occasional, off-line estimation of **K** would be possible for fast, optimal updating of the map at regular time intervals. As well as current AOD values, past values could be used to improve on FRK (i.e., smaller MSPEs); my IWGGMS-5 paper at Caltech on 6/25/08 addresses this problem

Epilogue: Still Thinking Statistically

- Regression depends as much on what X is chosen to regress on Y as what other variable was *not* chosen
- The variable $A =$ cloud microphysical properties (say), potentially influences the (regression) relationship between $Y =$ precipitation, and $X =$ aerosol

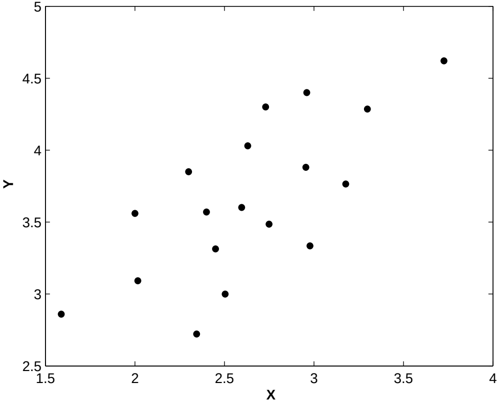
Stratify (on variable A)

$$\begin{aligned} \text{Stratum } 1 &: \quad \{\mathbf{s}_i : i \in A_1\} \\ \text{Stratum } 2 &: \quad \{\mathbf{s}_i : i \in A_2\} \\ \text{Stratum } 3 &: \quad \{\mathbf{s}_i : i \in A_3\} \\ A_1 \cup A_2 \cup A_3 &= \quad \{1, \ldots, n\} \end{aligned}$$

Summarize with a scatter plot that highlights the three different strata, and three statistical regression relationships. In region k ($k = 1, 2, 3$):

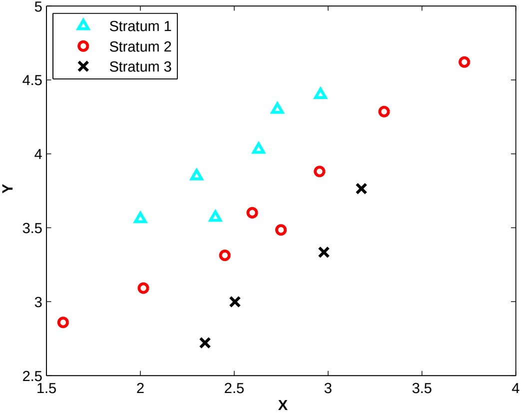
$$Y = \beta_{0,k}^{(A)} + \beta_{1,k}^{(A)} X + error$$

Scatterplot of Y versus X



$$Y = \beta_0 + \beta_1 X + error$$

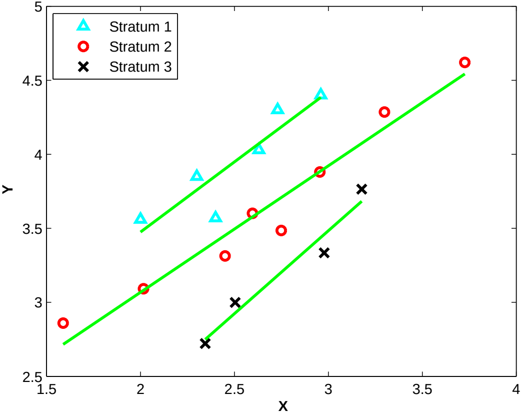
Stratification on Variable *A*



For 3 strata ($k = 1, 2, 3$)

$$Y = \beta_{0,k}^{(A)} + \beta_{1,k}^{(A)} X + error$$

Stratification on Variable A



$$\hat{\beta}_{1,1}^{(A)} > 0$$

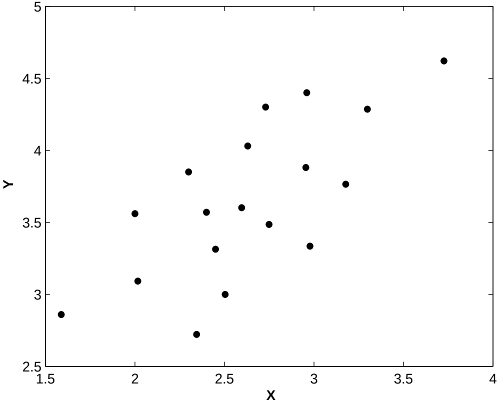
$$\hat{\beta}_{1,2}^{(A)} > 0$$

$$\hat{\beta}_{1,3}^{(A)} > 0$$

For 3 strata ($k = 1, 2, 3$)

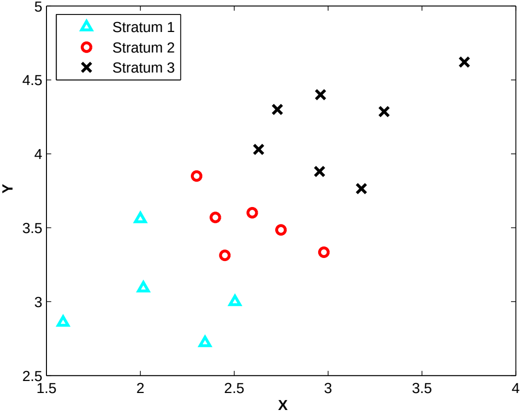
$$Y = \beta_{0,k}^{(A)} + \beta_{1,k}^{(A)} X + error$$

Scatterplot of Y versus X



$$Y = \beta_0 + \beta_1 X + error$$

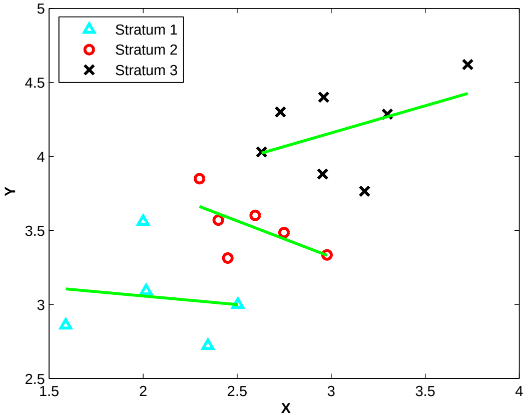
Stratification on Variable *B*



For 3 strata ($k = 1, 2, 3$)

$$Y = \beta_{0,k}^{(B)} + \beta_{1,k}^{(B)} X + error$$

Stratification on Variable B



$$\hat{\beta}^{(B)}_{1,1} < 0$$

$$\hat{\beta}^{(B)}_{1,2} < 0$$

$$\hat{\beta}^{(B)}_{1,3} > 0$$

For 3 strata ($k = 1, 2, 3$)

$$Y = \beta^{(B)}_{0,k} + \beta^{(B)}_{1,k} X + error$$

Aggregate (on variable A)

In Stratum k ($k = 1, \ldots, K$): Summarize with aggregated data $(\overline{X}_k, \overline{Y}_k)$

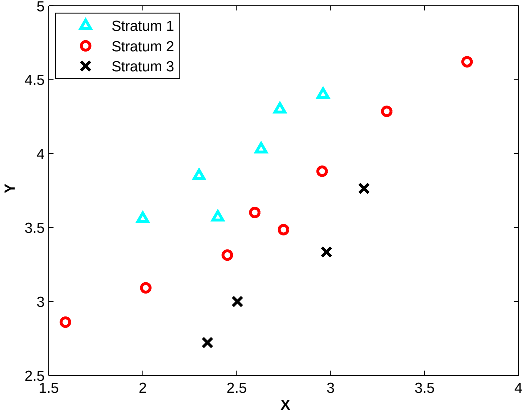
This is done all the time!

Derived (aggregated) data: $\{(\overline{X}_1, \overline{Y}_1), \ldots, (\overline{X}_K, \overline{Y}_K)\}$ for Stratum 1,. . . ,Stratum K on a given day.

Summarize with a scatter plot and a statistical regression relationship,

$$\overline{Y} = \alpha_0 + \alpha_1 \overline{X} + error$$

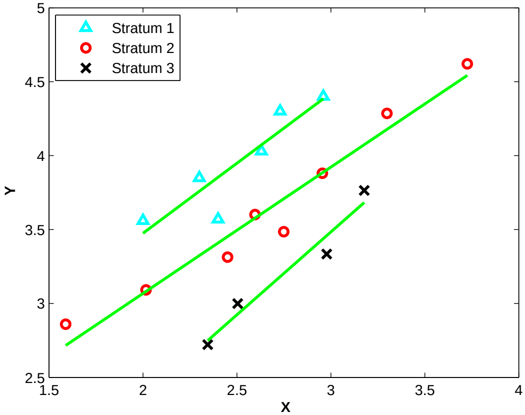
Stratification on Variable *A*



For 3 strata ($k = 1, 2, 3$)

$$Y = \beta_{0,k}^{(A)} + \beta_{1,k}^{(A)} X + \mathit{error}$$

Stratification on Variable A



$$\hat{\beta}^{(A)}_{1,1} > 0$$

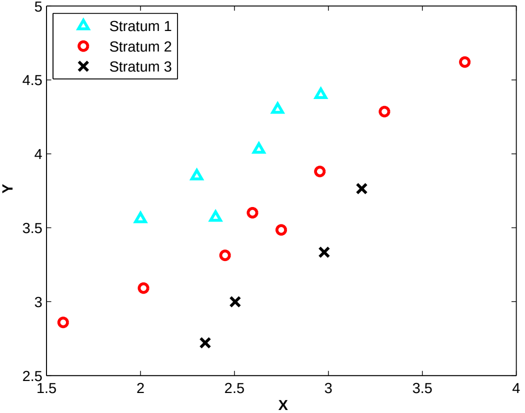
$$\hat{\beta}^{(A)}_{1,2} > 0$$

$$\hat{\beta}^{(A)}_{1,3} > 0$$

For 3 strata ($k = 1, 2, 3$)

$$Y = \beta^{(A)}_{0,k} + \beta^{(A)}_{1,k} X + error$$

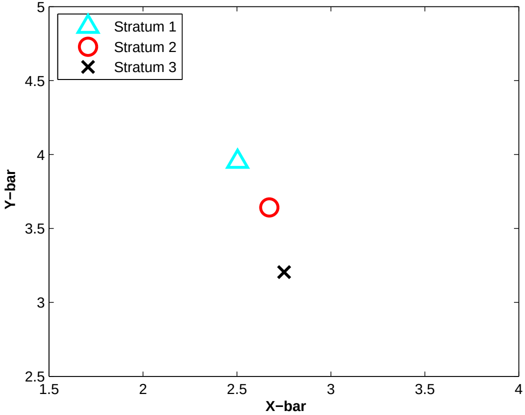
Stratification on Variable *A*



For 3 strata ($k = 1, 2, 3$)

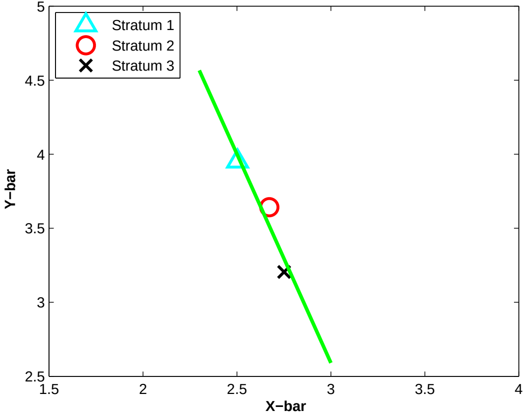
$$Y = \beta_{0,k}^{(A)} + \beta_{1,k}^{(A)} X + error$$

Aggregation Using Variable A



$$\overline{Y} = \alpha_0^{(A)} + \alpha_1^{(A)}\overline{X} + error$$

Aggregation Using Variable A



$$\hat{\alpha}_1^{(A)} < 0$$

$$\overline{Y} = \alpha_0^{(A)} + \alpha_1^{(A)}\overline{X} + error$$

Conclusions

- Regression fills gaps in noisy data with known mean squared prediction error
- Kriging (spatial regression) fills gaps in noisy spatial data with known kriging variance
- Very large spatial datasets can be kriged with a flexible family of spatial covariance functions
- Regression requires care when stratifying and aggregating. Solution: Build conditional probability models: $[Y \mid X, A]$

Conclusions

- Regression fills gaps in noisy data with known mean squared prediction error
- Kriging (spatial regression) fills gaps in noisy spatial data with known kriging variance
- Very large spatial datasets can be kriged with a flexible family of spatial covariance functions
- Regression requires care when stratifying and aggregating. Solution: Build conditional probability models: $[Y \mid X, A]$

Conclusions

- Regression fills gaps in noisy data with known mean squared prediction error
- Kriging (spatial regression) fills gaps in noisy spatial data with known kriging variance
- Very large spatial datasets can be kriged with a flexible family of spatial covariance functions
- Regression requires care when stratifying and aggregating. Solution: Build conditional probability models: $[Y|X, A]$

Conclusions

- Regression fills gaps in noisy data with known mean squared prediction error
- Kriging (spatial regression) fills gaps in noisy spatial data with known kriging variance
- Very large spatial datasets can be kriged with a flexible family of spatial covariance functions
- Regression requires care when stratifying and aggregating. Solution: Build conditional probability models: $[Y \mid X, A]$